StyloThai: A Scalable Framework For Stylometric Authorship Identification of Thai Documents

RAHEEM SARWAR, School of Information Science and Technology, VISTEC, Thailand
THANASARN PORTHAVEEPONG, School of Information Science and Technology, VISTEC, Thailand
ATTAPOL RUTHERFORD, Department of Linguistics at Faculty of Arts Chulalongkorn University, Thailand
THANAWIN RAKTHANMANON, Department of Computer Engineering, Kasetsart University, Thailand and School of Information Science and Technology, VISTEC, Thailand
SARANA NUTANONG*, School of Information Science and Technology, VISTEC, Thailand

Authorship identification helps to identify the *true* author of a given *anonymous* document from a *set of candidate authors*. The applications of this task can be found in several domains such as *law enforcement agencies and information retrieval*. These application domains are not limited to a specific *language, community or ethnicity*. However, most of the existing solutions are designed for English and a little attention has been paid to *Thai*. These existing solutions are not directly applicable to Thai due to the *linguistic* differences between these two languages. Moreover, the existing solution designed for Thai is *unable to* (i) handle outliers in the dataset; (ii) scale when the size of the candidate authors set increases; and (iii) perform well when the number of writing samples for each candidate author is low. We identify a stylometric feature space for the Thai authorship identification task. Based on our feature space, we present an authorship identification solution that uses *probabilistic k nearest neighbors'* classifier by transforming each document into a *collection of point sets*. Specifically, this document transformation allows us to (i) use set distance measures associated with outlier handling mechanism; (ii) capture stylistic variations within a document; and (iii) produce multiple predictions for a query document. We create a new Thai authorship identification corpus containing 547 documents from 200 authors, which is significantly larger than the corpus used by the existing study (an increase of 32 folds in terms of the number of candidate authors). The experimental results show that our solution can overcome the limitations of the existing solution and outperforms all competitors with an accuracy level of 91.02%. Moreover, we investigate the effectiveness of each *stylometric features category* with the help of an ablation study. We found that combining all categories of the stylometric features outperforms the other combinations. Finally, we cross-compare the feature spaces and classification methods of all solutions. We found that (i) our solution can scale as the number of candidate authors increases; (ii) our method outperforms all the competitors; and (iii) our feature space provides better performance than the feature space used by the existing study.

CCS Concepts: • **Information systems** → **Content analysis and feature selection**; • **Applied computing** → *Investigation techniques; Evidence collection, storage and analysis*; • **Computing Methodologies** → *Supervised learning by classification; Nearest neighbor algorithms.*

Additional Key Words and Phrases: Authorship analysis, stylometry, similarity search, Thai authorship identification

*Corresponding Author

Authors' addresses: Raheem Sarwar, School of Information Science and Technology, VISTEC, Wangchan Valley 555 Moo 1 Payupnai, Wangchan, Rayong, Thailand, 21210, raheem.s@vistec.ac.th; Thanasarn Porthaveepong, School of Information Science and Technology, VISTEC, Wangchan Valley 555 Moo 1 Payupnai, Wangchan, Rayong, Thailand, 21210, Thanasarn.p@vistec.ac.th; Attapol Rutherford, Department of Linguistics at Faculty of Arts Chulalongkorn University, Bangkok, Thailand, attapolrutherford@gmail.com; Thanawin Rakthanmanon, Department of Computer Engineering, Kasetsart University, Thailand , School of Information Science and Technology, VISTEC, Wangchan Valley 555 Moo 1 Payupnai, Wangchan, Rayong, Thailand, 21210, thanawin.r@ku.ac.th; Sarana Nutanong, School of Information Science and Technology, VISTEC, Wangchan Valley 555 Moo 1 Payupnai, Wangchan, Rayong, Thailand, 21210, snutanon@vistec.ac.th.

1 Introduction

Stylometry is the science of measuring the *writing style* of an author [1, 7, 14]. *Stylometry* relies on the observation that each author has a unique writing style which can help to differentiate among the documents written by different authors [29]. This concept is known as the *authorial fingerprint* [23, 29, 32]. Coulthard [4] describes that, every author has his/her own form of the language, which is known as *idiolect*. An authors' *idiolect* manifests itself *distinctive* and *cumulatively unique rule-governed* choices for the written communication. Specifically, every author has stored a large set of vocabulary built up over many years. The vocabulary-set of an author may differ *considerably* or *slightly* from the vocabulary-set of all other authors. The difference among the vocabulary-sets of authors occurs in terms of the *stored vocabulary items*, *passive vocabulary items* and most importantly *their preferences for selecting and combining these items* for written communication. These differences can help identify authors. One prominent task performed by using stylometry is authorship identification, which identifies the original author of a given *anonymous* document, and is formally defined as follows.

Definition 1.1 (Authorship Identification). "Given an anonymous/disputed text x, a set of candidate authors Y, and their writing samples X, identify the most likely author of x in Y by analyzing the writing samples in X and comparing them with x [23]. "

Applications of the authorship identification task span across several areas such as *intelligence agencies work*, where authorship identification can help linking the *intercepted* messages to known enemies [1, 21, 29]; *criminal law*, where authorship identification can help identifying the true author of harassing letters and ransom notes [29]; and *plagiarism detection*, where an authorship identification solution can help identifying the true authors of student submissions [23]. Moreover, these days, managing large text repositories has become a major challenge and it has received significant attention by researchers, from several areas, such as web information management [25], natural language processing [34], and information retrieval [38].

The application domains of authorship identification are not limited to a specific language, community or ethnicity. Thai is a member of the Kra-dai languages family. The Kra-dai languages[1] include Thai, Lao, the tonal languages spoken in Southeast Asia, Northeast India and Southern China. More than 90 million people speak Kra-dai languages and *Thai* is the most widely spoken Kra-dai language. However, most of the existing authorship identification solutions are designed for English [15, 16, 18]. These solutions are not directly applicable to Thai due to linguistic differences between English and Thai. For example, unlike English [15, 16, 18], (i) Thai has 44 consonants and 4 tone marks; (ii) Thai has 18 vowel symbols that create many compound vowels, and few special symbols; (iii) Thai text samples do not have word/sentence boundaries; and (iv) the first person pronounce in Thai can be gender-specific and gender-neutral (i.e. men tends to use the former and women tends to use the later) [18]. These characteristics of Thai makes the stylometric features extraction process noisier in comparison to English, and requires a solution associated with outlier handling techniques (see Section 4.1 for more details). To the best of our knowledge, only one solution [18] has been proposed to perform authorship identification for Thai, and we call it SVM-CWPEST for short (see Section 2.1.3 for more details about this existing solution). However, there are several limitations of this solution which can be described as follows.

Limitations of Existing Study.

(1) ***Low Accuracy of Language Processing Tools.*** The aforementioned unique characteristics of Thai such as, Thai text does not contain any word/sentence boundary, makes it harder

[1] Available from https://en.wikipedia.org/wiki/Kra-Dai_languages. Retrieved 09/07/2019

for *Thai language processing tools* such as TLTK[2] to yield a high accuracy [5, 6, 16, 18]. Consequently, the stylometric features extraction process for Thai is noisier in comparison to English. However, the existing solution (SVM-CWPEST) [18] is not associated with any noise handling mechanism. *We aim at designing an authorship identification solution for Thai that can mitigate the effect of outliers in the dataset to improve the authorship identification accuracy.*

(2) ***Small Number of Writing Samples Per Author.*** The existing solution (SVM-CWPEST) [18] is unable to perform well when the average number of writing samples for a candidate author is between 2 and 3 (see Section 5.2 for more details). Using a large number of different documents (i.e., writing samples) helps better capture the *stylistic variation information* of an author. For example, the existing study [18] uses 25 writing samples for each candidate author. However, such a large number of writing samples may not be available for each candidate author in real-world scenarios. *Thus, we aim at designing an authorship identification solution for Thai which can perform well (i.e., achieve more than 90% accuracy) when the average number of writing samples for each candidate author is low, i.e., less than 3.*

(3) ***Large Size of Candidate Author Set.*** The existing solution (SVM-CWPEST) [18] drastically drops the accuracy when the size of candidate author set increases (see Section 5.2 for more details). Moreover, the existing study [18] is limited to 6 candidate authors only. However, in a real-world scenario, such as plagiarism detection in student submissions, there can be hundreds of candidate authors. *We aim at designing an authorship identification solution for Thai which can handle large number of candidate authors.*

In this investigation we identify a *stylometric feature space* (LSS) to perform authorship identification on Thai. Specifically, our feature space (LSS) consists of 46 stylometric features which can be organized into three main categories including: 27 lexical features (L), 17 syntactic features (S) and 2 structural features (S). These features are explained in Section 4.1. Our feature space (LSS) is better than the feature space (CWPEST) used in existing work [18]. This is because, unlike CWPEST, the LSS feature space contains syntactic features (i.e., part-of-speech (POS) based features), which can play an important role in distinguishing between documents written by different authors [14, 23, 32] (see Section 5.2 for experimental results).

Research Questions. In addition to addressing the aforementioned limitations of the existing study [18], we answer the following research questions in this paper.

- ***Research Question 1.*** Recall that, unlike the feature space (CWPEST) used by existing work [18] which does not contain syntactic features, our stylometric feature space (LSS) contains the *syntactic features* in addition to the lexical and structural features. Thus, we investigate how important it is to use syntactic features for Thai authorship identification process? In addition to this, we also investigate the importance of each category of the stylometric features to perform Thai authorship identification with the help of an ablation study.
- ***Research Question 2.*** How important it is to use all three categories of the stylometric features in the authorship identification process?
- ***Research Question 3.*** How important it is to use set similarity measures associated with outlier handling mechanisms in comparison to the standard set similarity measure (i.e., without outlier handling mechanism), in Thai authorship identification process. The set similarity measures are discussed in Section 4.2.

As for the classification method, we adopt the *probabilistic k nearest neighbors* classifier to perform scalable authorship identification with limited number of writing samples per candidate

[2] https://pypi.org/project/tltk/

author [11]. However, the PkNN is sensitive to noise in the dataset [11]. To address this issue, we use a document transformation model that relies on set similarity search [33] such that the stylistic variations between the text samples can be computed as a set distance [12]. By using a corpus of 547 Thai documents from 200 authors, which is significantly larger than existing study (an increase of 32 folds in terms of the number of candidate authors), we perform experimental studies to show that our solution can (i) mitigate the effect of outliers in the dataset; (ii) handle a large number of candidate authors; (iii) perform well when the number of writing samples per candidate author is low; and (v) achieve the accuracy level of 91.02% which is higher than all competitors.

Summary of Our Contributions. The contribution of this work includes:

(1) We formulate an effective stylometric features space (LSS) for Thai authorship identification task . Based on LSS, we present an authorship identification solution for Thai that can overcome the limitations of existing study and achieve the accuracy level of 91.02%, which is higher than all competitors.
(2) We create a new significantly larger Thai authorship identification corpus than existing study (i.e., an increase of 32 folds in terms of number of candidate authors);
(3) We summarize the findings of our studies here to compare the performance our solution against (i) SVM-CWPEST, the only existing authorship identification solution for Thai; and (ii) four extensively used classifiers in authorship identification studies in different settings.

The rest of the paper is organized as follows. Section 2 reviews existing studies on authorship identification. Section 3 illustrates our corpus. Section 4 describes our solution. Section 5 presents the experimental results. Section 6 contains the concluding remarks.

2 Literature Review

Authorship identification is generally performed in two steps. The first step is related to the stylometric features extraction from the true documents of the candidate authors. The *stylometric features* are the writing style markers that can help distinguish among the documents from different authors. Stylometric features can be organized into three main categories, namely, lexical features, syntactic features, and structural features [20, 23, 24].

(1) The *lexical features* are the statistical measures of character-based and word-based lexical variations in a document, such as, vocabulary richness [29], and word length distributions [23].
(2) The *part-of-speech (POS)* tags and *function words* are the examples of the *syntactic features* [20].
(3) The *structural features* are associated with the organization of the document, such as average number of words in a sentence or a paragraph [23].

The second step is related to learning a classification model to predict the true author of the anonymous document.

2.1 Authorship Identification Methods

2.1.1 Deep Learning Based Methods to Authorship Identification. Recently, the deep learning methods have received a significant attention by researchers. Specifically, the deep learning methods do not require *manual features engineering* which makes them more effective over traditional techniques. This is due to the fact that a right set of features is required to achieve state-of-the-art accuracy [9, 13, 22, 27, 35, 37]. Nevertheless, a tremendous amount of data is required to train the deep learning models. That is, in a Convolutional Neural Network (CNN), the implicit *data representation* is learned in hidden layers, and based on this learned data representation the classification is performed at the output layer. Given the huge amount of training data, the learned data representation is better in comparison to hand-crafted features and provides the better accuracy.

Several existing studies focused on English used deep learning to perform the authorship identification task. For example, Solorio [36] performed authorship identification using a three-layer CNN model based on character bi-grams. They reported an accuracy level of 76.1% using a corpus written by 50 authors where each author has 1,000 samples. Moreover, Ge [8] performed authorship identification using the *feedforward neural networks*. The authors from [8] reported 95% accuracy and noted that this task is too easy to perform due to the availability of huge training data.
Comparison with Our Work. The deep learning methods may achieve high accuracy for authorship identification task only when a large amount of training data is available. However, in this investigation, we aim at designing an authorship identification solution for Thai which can perform well in *data-poor conditions* where the average number of writing samples for each candidate author is between 2 and 3.

2.1.2 Machine Learning Methods to Authorship Identification. The well-known machine learning methods for authorship identification includes random forests (RF), decision trees (DT), naïve bayes (NB), and support vector machines (SVM) [1, 18, 29, 33]. In this work, we compare the accuracy of our method against these well-known extensively used methods by varying the size of the candidate author set. In addition to directly comparing our solution against these competitors, we cross-compare the feature spaces and methods by formulating different solutions (see Section 5.2 for more details). The implementation details of these methods are given in Table 1. Specifically, we use the WEKA's implementation as given in Table 1. Among these methods, the LibSVM is not available directly in WEKA and we included it manually.

Table 1. Implementations of the Classification Algorithms and Their Parameters (* Available under WEKA.Classifiers)

Method	Implementation	Parameters changed from the Default
Support Vector Machines (SVM)	*.functions.LibSVM	–
Naïve Bayes (NB)	*.bayes.NaïveBayes	kernel: Radial Basis
Decision Trees (DT)	*.trees.J48	–
Random Forests (RF)	*.trees.RandomForests	–

2.1.3 Thai Authorship Identification (SVM-CWPEST). We note that most existing solutions are designed for English. To the best of our knowledge, there is only one study which is focused on authorship identification of Thai text. This study is performed on a corpus from 6 authors where each author has 25 text samples [18]. This study extracts 53 features called CWPEST from each writing sample and apply the *Support Vector Machines (SVM)* and Decision Trees (Weka's J48) classifiers to predict the true author of the anonymous text, and shows that SVM yields better accuracy. We call this Thai authorship identification solution SVM-CWPEST for short.
Comparison With Our Method. Note that unlike our method that represents each document as a collection of point sets, the SVM-CWPEST method represents each document as *one single data point* in a multidimensional space (see Section 4 for more details). As a result, the SVM-CWPEST method is *unable to* (i) capture the writing style variations within the same document; (ii) produce multiple predictions for the same document; and (iii) apply set distance measures to handle outliers in the dataset. Moreover, unlike the feature space used by existing method (CWPEST), our feature space contains POS-based features in addition to the lexical and structural features. Moreover SVM-CWPEST was applied on short text samples. The applications of *short-text*

authorship identification can be found in social media domain, such as, author identification of controversial posts by virtual identities on social media, authorship analysis on Facebook posts, Twitter status, chat conversations and Short Message Service (SMS) messages [1]. The applications of long-text authorship identification can be found in several areas associated with managing large text repositories and plagiarism detection in student theses. Specifically, retrieving and categorizing documents *with respect to their authors* and plagiarism detection have been receiving significant attention by researchers in several areas such as *natural language processing* [2, 3, 34], *web information management* [25] and *information retrieval* [10, 26, 28, 30, 31, 38].

3 Data Collection

There are two main issues associated with authorship identification corpora: (i) the number of publicly available corpora is limited; and (ii) the size of the publicly available corpora is small in terms of the number of candidate authors. To the best of our knowledge, there is no benchmark corpora available for Thai authorship identification task. In order to perform experiments, we created a new Thai authorship identification corpus extracted from an online *Dek-D*[3] repository. Our scraper is written in Python and extracts the data in two steps: (i) retrieve all the URLs of each author; and (ii) based on the retrieved URLs, extract the documents of each author from the website. Our corpus contains 547 Thai documents from 200 authors where the average length of documents is 25,334 tokens. Moreover, our corpus is significantly larger than existing study [18] (i.e., an increase of 32 folds in terms of the number of candidate authors). Furthermore, on average there are 2.73 samples per class (author) which is more realistic scenario where a large number of writing samples per author may not be available.

4 Methodology

We explain our solution with the help of Fig. 1. Our solution consists of four main parts including (i) preprocessing, (ii) set similarity search, (iii) probabilistic k nearest neighbor classification, and (iv) prediction aggregation.

4.1 Preprocessing

The preprocessing part of our solution transforms each document into a *collection of fragments* (i.e., *collection of point sets*) using a three steps process [33]: (i) partition each document into fixed size fragments; (ii) partition each fragment obtained from the first step into fixed size chunks[4]; and (iii) extract the 46 stylometric features (i.e., writing style markers) from each chunk. To obtain reliable stylometric statistics from each fragment and the chunk, we fix their sizes to 7,000 and 700 tokens[5], respectively. As a result, each chunk is transformed into a point, each fragment is transformed into a *point set* and each document is transformed into a *collection of point sets* in 46-dimensional space. There are three advantages of transforming each document into a *collection of point sets*. (1) We can compute the *stylistic variations* between text samples as a *set distance*. Specifically, we can use those set distance measures which are capable of mitigating the effect of outliers in the data such as *partial Hausdorff distance* [12]. (2) We can capture the *stylistic variation* of an author within a document. This is because, each authorship identification prediction is produced by multiple points rather than one single point. (3) We can produce multiple predictions for a query document, which allows us to use only the most certain γ% predictions of a query document for the prediction aggregation process, as explained latter in this section. Once we complete the feature

[3] https://www.dek-d.com

[4] A chunk is a collection of Tokens

[5] token is the content of text separated by white space character.

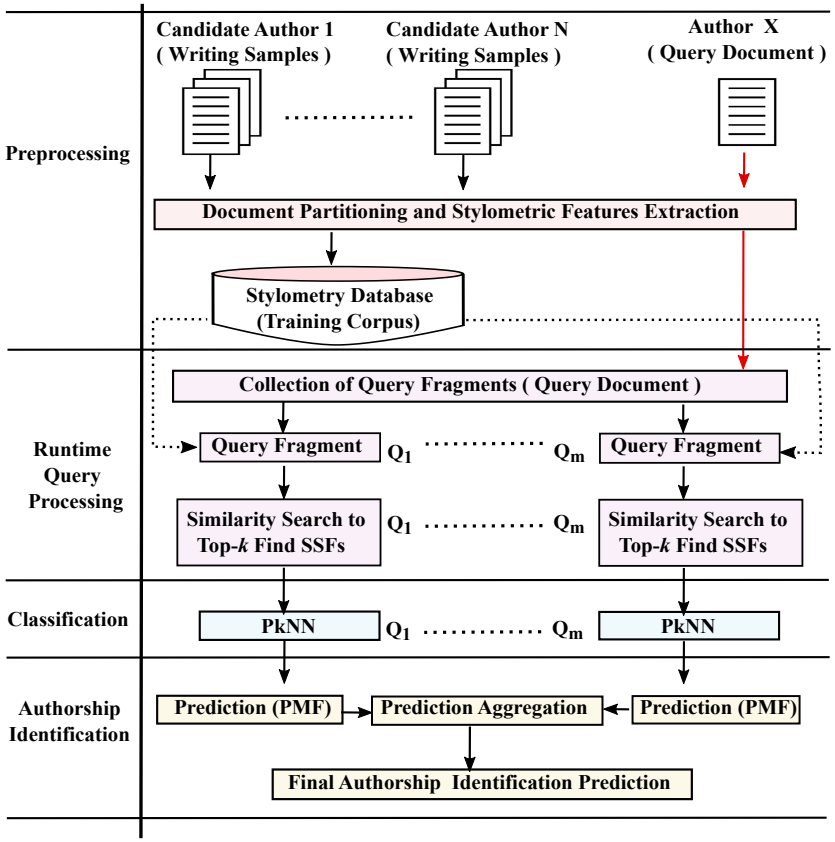

Fig. 1. The Overview of StyloThai Framework [33]

extraction process, we store the feature values into the stylometry database. Our 46-dimensional stylometric feature space can be organized into three categories as shown in Table 3: (i) lexical features (from # 1 to # 27), (ii) syntactic features (from # 28 to # 44), and (iii) structural features (from # 45 to # 46). Note that, in our stylometric feature space given in Table 3, 32 out of 46 features (i.e., except the character-based features from # 14 to # 27) require the word tokenization and sentence identification, which is a challenging task to perform due to the reason that there are no word/sentence boundaries in Thai. These characteristics of Thai make it harder for *Thai language processing tools* such as TLTK[6] to yield a high accuracy [5, 6, 16, 18]. Consequently, the stylometric features extraction process for Thai is noisier in comparison to English. Hence, our Thai authorship identification solution is associated with outlier handling mechanisms as discussed later in this section.

In Table 3, N represents the count of words, V represents the count of distinct words, V_i represents the count of words that occur i times, and C represents the total number of characters. As for

[6]'https://pypi.org/project/tltk/'

the lexical features, we identified 13 word-based (from #1 to #13) and 15 character-based (from # 14 to # 27) stylometric features and computed them from each chunk. These 13 word-based lexical variation features can be considered as language-independent [29]. For the rest of the 15 character-based lexical features, 5 features are specific to Thai, i.e., the feature # 16, 17, 21, 23 and 24. As for the syntactic features, we identified 17 features (from # 28 to # 44) based on the relative frequency of parts-of-speech categories and computed them from each chunk. Finally, we identified 2 structural features (from # 45 to # 46) based on the text-organization, such as, average number of words per sentence and total number of sentences in a chunk. The word tokenization is performed using DeepCut[7]. The rest of the features are calculated using Thai Language Toolkit (TLTK)[8]. We provide an example of Thai text sample in Table 2 and show the computed stylometric feature values in Table 3. We measure the effectiveness of each features category for authorship identification process (see Section 5.2 for experimental results).

Table 2. An Example of Stylometric Features Extraction from a Chunk

Thai
(๓๐ พ.ค. ๒๕๖๒) 07.00 น. @ศูนย์วิทยุ 191 ได้รับแจ้งว่าพบวัตถุต้องสงสัยเป็นระเบิดพร้อมปืนM60จำนวนมากฝังอยู่ใต้ต้นข่อย หลังจากเขาได้ยินเสียงดังตูม บริเวณบ้านไม่มีเลขที่ ในหมู่บ้านเฉลิมพระเกียรติ จึงรีบรุดไปตรวจสอบบริเวณดังกล่าว พร้อมกับเจ้าหน้าที่ชุดเก็บกู้
English (Word Based Translation)
(30 MAY 2019) 7 o'clock @ Radio Center 191 was informed that the founded suspected object was a bomb and many M60 guns buried under Tooth brush tree. After he heard loud noises around the unnumbered house in the Chaloem Phra Kiat village, thus hurried to check the area along either the EOD staff.

4.2 Set Similarity Search

While processing a given query document (Q), we first apply the preprocessing step of our solution on Q which transforms it into a *collection of point sets*. We then execute an independent set similarity query for each query fragment (Q) in Q to retrieve top-k SSFs from the corpus. Note that, we execute an individual *set similarity query* for each Q in Q, i.e., if a query document results in m query fragments (point sets), we execute m independent set similarity queries (see Figure 1). While retrieving the top-k SSFs, we tried three set similarity measures including, (i) *standard Hausdorff Distance (SHD)*, (ii) *partial Hausdorff Distance (PHD)* [12] and *modified Hausdorff Distance (MHD)* [17] as a proximity measure between two point sets. The SHD between two points sets Q and F can be calculated as:

$$h(Q, F) = \max_{q_i \in Q} \min_{f_j \in F} d(q_i - f_j).$$

That is, SHD can be calculated by: (i) ranking all data points in a query fragment Q in accordance with the *minimum distance* to the fragment F; and (ii) *selecting the maximum of the minimum distances*. Researchers have argued that SHD is sensitive to the noise in the data [12, 17]. To mitigate the noise (outlier) sensitivity issue associated with SHD, researchers formulated two variants of

[7]'https://github.com/rkcosmos/deepcut'
[8]'https://pypi.org/project/tltk/'

Table 3. List of Stylometric Features (N represents the count of words, V represents the count of distinct words, V_i represents the count of words that occur i times, and C represents the Total number of characters)

Lexical Features			
Stylometric Features	Values	Stylometric Features	Values
1. **N:** Total # words	72	2. **V:** Total # distinct words	59
3. Average word length	3.53	4. S.D. of word lengths	22.18
5. $\frac{V}{N}$	0.82	6. $VR(K) = \frac{10^4(\sum i^2 V_i - N)}{N^2}$	223.7654321
7. $VR(R) = \frac{V}{\sqrt{N}}$	6.95	8. $VR(C) = \frac{\log V}{\log N}$	0.95
9. $VR(H) = \frac{(100 \log N)}{(1-V_1)/V}$	-467.26	10. $VR(S) = \frac{V_2}{V}$	0.050847458
11. $VR(k) = \frac{logV}{\log(\log N)}$	2.81	12. $VR(LN) = \frac{(1-V^2)}{V^2(\log N)}$	-0.23
13. Entropy of word freq. ditri.	872.03	14. **C:** Total # chars	254
15. Freq. of alpha chars	1	16. Freq. of Thai chars	219
17. Freq. of Thai numeric chars	6	18. Freq. of Arabic numeric chars	9
19. Freq. of special chars	7	20. Freq. of white spaces	12
21. Freq. of vowel and tone marks	86	22. $\frac{\text{Freq. of Alpha char}}{\text{C}}$	0.0039
23. $\frac{\text{Freq. of Thai char}}{\text{C}}$	0.85	24. $\frac{\text{Freq. of Thai numeric char}}{\text{C}}$	0.024
25. $\frac{\text{Freq. of Arabic numeric char}}{\text{C}}$	0.035	26. $\frac{\text{Freq. of Special char}}{\text{C}}$	0.028
27. $\frac{\text{Freq. of White spaces}}{\text{C}}$	0.047		
Syntactic Features			
Stylometric Features	Values	Stylometric Features	Values
28. $\frac{\text{Freq. of adjectives}}{\text{N}}$	0.014	29. $\frac{\text{Freq. of adpositions}}{\text{N}}$	0.042
30. $\frac{\text{Freq. of adverbs}}{\text{N}}$	0.028	31. $\frac{\text{Freq. of auxiliarys}}{\text{N}}$	0.014
32. $\frac{\text{Freq. of coordinating conjunctions}}{\text{N}}$	0.014	33. $\frac{\text{Freq. of determiners}}{\text{N}}$	0.014
34. $\frac{\text{Freq. of interjections}}{\text{N}}$	0.014	35. $\frac{\text{Freq. of nouns}}{\text{N}}$	0.264
36. $\frac{\text{Freq. of numerals}}{\text{N}}$	0.042	37. $\frac{\text{Freq. of particles}}{\text{N}}$	0.014
38. $\frac{\text{Freq. of pronouns}}{\text{N}}$	0.013	39. $\frac{\text{Freq. of proper nouns}}{\text{N}}$	0.014
40. $\frac{\text{Freq. of punctuation}}{\text{N}}$	0.19	41. $\frac{\text{Freq. of subconjunction}}{\text{N}}$	0.042
42. $\frac{\text{Freq. of symbols}}{\text{N}}$	0.013	43. $\frac{\text{Freq. of verbs}}{\text{N}}$	0.181
44. $\frac{\text{Freq. of other POS}}{\text{N}}$	0.014		
Structural Features			
Stylometric Features	Values	Stylometric Features	Values
45. Total number of sentence	11	46. Avg. #words per sentence	0.153

SHD: *modified Hausdorff distance (MHD)* [17] and *partial Hausdorff distance (PHD)* [12]. Specifically, the MHD and PHD measures average out the effect of the outlier over the minimum distances falling into a specified range i.e., $(50\%, 100\%]$ (for MHD, the second parameter values is always 100%). The experimental results regarding set distance measures are reported in Section 5.2.

4.3 Probabilistic k Nearest Neighbors Classification (PkNN)

We apply PkNN [11] to the retrieved top-k SSFs to make a probabilistic prediction for each query fragment in a query document. Unlike simple kNN classifier where the output is one single class (author), the PkNN classifiers produces a probability mass function (PMF) over all classes (candidate authors) associated to the retrieved SSFs. We apply the PkNN [11] that utilizes the distance values of the k nearest neighbors (SSFs in this case) to weight the distribution of the probability. An exponential function is used to smooth the distance-probability mapping [29]. The advantages of using PkNN [11] over other classifiers can be summarized as follows. A little or no training is required for classification [19]. Consequently, there is no information loss associated with generalization [11, 29]. This classifier is capable of performing classification with a limited set of samples [29]. Moreover, it allowed us to apply set distance measures, capable of mitigating the effect of outliers in the dataset [12].

4.4 Prediction Aggregation

The final step of our solution is to merge all the *fragment probabilistic predictions* such that *one single authorship identification prediction* can be produced for the entire $\boldsymbol{Q}$ [33]. In order to do so, one can simply compute the average of all fragment probabilistic predictions. However, all the fragment probabilistic predictions (one for each Q) of a query document $\boldsymbol{Q}$ are not equally useful, i.e., there can be highly uncertain predictions and including them into the prediction aggregation process may damage the overall accuracy [33]. At this stage, we apply *entropy* as an uncertainty measure to find the uncertain fragment predictions and eliminate them from the prediction aggregation process [33]. The final probabilistic prediction of the entire $\boldsymbol{Q}$ is computed as the average PMF of most certain γ% prediction. An example of this process is given in the Table 4. Assume that the value of γ is 50. The top 50% most certain predictions belong to Q_2 and Q_3 as indicated with $*$ (with the low entropy values). The final prediction of the entire query document $\boldsymbol{Q}$ is calculated as the average PMF of Q_2 and Q_3.

Table 4. An Example of Prediction Aggregation Process (*Top most certain γ50% predictions)

Query Fragment (Q)	Query Fragment Prediction (PMF)			Entropy
Q_1	[$Author\ A$: 0.33,	$Author\ B$: 0.34,	$Author\ C$: 0.33]	1.5848
Q_2^*	[$Author\ A$: 0.36,	$Author\ B$: 0.32,	$Author\ C$: 0.32]	1.5827
Q_3^*	[$Author\ A$: 0.32,	$Author\ B$: 0.35,	$Author\ C$: 0.33]	1.5840
Q_4	[$Author\ A$: 0.33,	$Author\ B$: 0.34,	$Author\ C$: 0.33]	1.5848
Final Prediction	[$Author\ A$: 0.34,	$Author\ B$: 0.335,	$Author\ C$: 0.325]	–

5 Performance Evaluation

5.1 Experimental Setup

Evaluation Measures. Recall that we represent each document as a collection of fragments. Hence, we compute the authorship identification accuracy at two levels, (i) fragment level; and (ii) document level as follows:

- *Fragment Accuracy:* "A fragment authorship prediction is considered correct if the true author of the query document is identified as the most likely author".

- *Document Accuracy:* "An aggregated final authorship prediction of the query document is considered correct if the true author of the query document is identified as the most likely author".

Parameter Setting. Although we have not shown here, we tried different values for each parameter, and the parameter values given in Table 5 resulted in the best accuracy. The k value denotes the number of closest stylistically similar fragments identified as a result of set similarity search query, to use for PkNN. The values, (50%,100%] and (50%,75%], denote the MHD and PHD ranges, respectively. The L and l denote the sizes of each fragment and each chunk respectively. The γ denotes the percentage of predictions that we consider for the prediction aggregation process illustrated in section 4.4.

Table 5. Default Parameter values of our method

k	MHD	PHD	L	l	γ
5	(50%, 100%]	(50%, 75%]	7,000 tokens	700 tokens	90%

Evaluation Strategy. To evaluate the accuracy of all methods in this investigation, we use 5-fold cross validation. Recall that, as for our method, each document is represented as a collection of fragments. To avoid test-train set contamination in the evaluation process of our method we ensure that when a document is used for testing it is purely used for testing.

5.2 Experimental Results

In this section we report results from our experimental studies. Note that, all experiments are performed using corpora containing limited number of writing samples per candidate author i.e., between 2 and 3.

An Ablation Study of Different Features (Effect of Feature Types). This study provides the answers to the first two questions mentioned in Introduction Section. As can be seen from Table 6 that (i) including syntactic features into lexical + structural increases the authorship identification accuracy from 61.21% to 91.02%; and (ii) combining all categories of stylometric features (i.e., lexical + syntactic + structural) outperforms the other combinations (i.e., (a) lexical + structural, (b) syntactic + structural, and (c) lexical + syntactic) . These results indicate that the stylometric information captured by different features categories is complementary and orthogonal. Consequently, combining all feature categories improves the performance of authorship identification process. Hence, we confine rest of the experimental studies to combined categories of the stylometric features only (i.e., lexical + syntactic + structural).

Table 6. StyloThai Document Accuracy: Effect of Feature Types

Lexical	Syntactic	Structural	Accuracy
✓		✓	61.21%
	✓	✓	70.23%
✓	✓		79.83%
✓	✓	✓	**91.02%**

Effect of Set Distance Measures and Prediction Aggregation Process. In this study, by using a corpus containing 547 Thai document from 200 authors, which significantly larger than the corpus

used by existing study (an increase of 32 folds in terms of the number of candidate authors), we show that our method can (i) mitigate the effect of outliers in the dataset; (ii) handle a large number of candidate authors; and (iii) perform well in extreme data poor condition. The experimental results obtained using our method are shown in Table 7. As mentioned in Section 4.2 that, unlike MHD and PHD distance measures, SHD is not associated with outlier handling mechanism. The fact that SHD is significantly outperformed by MHD and PHD shows that our dataset in fact has noise (outliers) to be handled. Moreover, the results show that PHD has better outlier handling mechanism than MHD. Due to the obvious accuracy gaps, we only adopt the PHD measure in rest of the experimental studies. Moreover, the experimental results show that instead of using 100% fragment predictions in the prediction aggregation process, using 90% most certain fragment-predictions yields better accuracy. Due to the obvious accuracy gap between fragment and the document accuracies, we only report document accuracy ($\gamma = 90\%$) in rest of the experimental studies.

Table 7. Proposed Method Only (StyloThai): The effect of set distance measures and γ value

	Accuracy				
Distance Measure	Fragment	Document ($\gamma = 50\%$)	Document ($\gamma = 70\%$)	**Document ($\gamma = 90\%$)**	Document ($\gamma = 100\%$)
SHD	73.22%	78.41%	79.09%	80.12%	78.34%
MHD	84.55%	88.66%	90.21%	90.43%	89.03%
PHD	**85.89%**	89.15%	90.47%	**91.02%**	89.14%

Effect of the Candidate Author Set Size. In this study, we provide the performance comparison between our solution and the competitive solutions by varying the size of the candidate author set from 50 to 200. In addition to directly comparing our solution against the competitors, we cross-compare the feature spaces and methods by formulating the following solutions.

(1) ***StyloThai-LSS :*** *Our feature space (LSS) applied to our method* (StyloThai) *[proposed solution].*
(2) **StyloThai-CWPEST :** Feature space used by existing study [18] (CWPEST) applied to *StyloThai* .
(3) **SVM-LSS :** LSS applied to the *support vector machines (SVM)* method.
(4) ***SVM-CWPEST :*** *CWPEST applied to* support vector machines (SVM) *[existing solution for Thai [18]].*
(5) **RF-LSS :** LSS applied to the *random forests (RF)* method.
(6) **RF-CWPEST :** CWPEST applied to the *random forests (RF).*
(7) **NB-LSS :** LSS applied to the *naïve bayes (NB)* method.
(8) **NB-CWPEST :** CWPEST applied to the *naïve bayes (NB)* method.
(9) **DT-LSS :** LSS applied to the *decision trees (DT)* method.
(10) **DT-CWPEST :** CWPEST applied to the *decision trees (DT)* method.

The experimental results given in Table 8 show that (i) our solution (StyloThai-LSS) can scale as the number of candidate authors increases; (ii) our method (StyloThai) outperforms all the competitors; and (iii) our feature space (LSS) provides better performance than the competitive feature space (CWPEST). Moreover, regardless of the feature space, there is a significant accuracy gap between our method (StyloThai) and other methods which are not associated with outlier handling mechanisms, i.e., SVM, RF, BN and DT.

Table 8. Document Accuracy: Effect of Candidate Author Set Size

	The effect of Number of Candidate Authors			
Method	50	100	150	200
StyloThai-LSS [Our Method]	**92.05%**	**92.13%**	**91.49%**	**91.02%**
StyloThai-CWPEST	77.67%	72.36%	69.04%	62.47%
SVM-LSS	44.63%	34.47%	25.42%	18.58%
SVM-CWPEST (Competitive Method)	33.84%	21.61%	12.91%	08.97%
RF-LSS	39.27%	25.39%	19.93%	17.34%
RF-CWPEST	34.91%	20.73%	11.82%	07.84%
NB-LSS	36.43%	24.29%	17.24%	15.97%
NB-CWPEST	29.78%	22.43%	13.79%	09.69%
DT-LSS	35.09%	21.96%	17.56%	16.88%
DT-CWPEST	27.05%	16.44%	14.75%	10.35%

6 Conclusions

This paper presents a scalable solution for authorship identification of Thai documents. The existing solutions designed for English are not directly applicable to Thai due to the linguistic differences between them. Moreover, the existing solution designed for Thai is (i) not associated with any outlier handling mechanism; (ii) unable to scale when the size of the candidate authors set increases; and (iii) cannot perform well in data-poor conditions. By using a corpus of 547 documents written in Thai from 200 authors, which is significantly larger than the corpus used by the existing study, we perform extensive experimental studies to show that our solution can (i) mitigate the effect of outliers in the dataset; (ii) handle a large number of candidate authors in extreme data-poor conditions; and (iii) achieve the accuracy level of 91.02% which is significantly higher than all competitors. In addition to addressing the aforementioned limitations of the existing study, we answer the following three research questions in this paper. (i) How important it is to use syntactic stylometric features in Thai authorship identification process? (ii) How important it is to use all three categories of stylometric features in authorship identification process? (iii) How important it is to use set similarity measures associated with outlier handling mechanisms in comparison to the standard set similarity measure (i.e., without outlier handling mechanism), in Thai authorship identification process. We found that (1) including the syntactic features into lexical + structural features increases the authorship identification accuracy from 61.21% to 91.02%; (2) combining all categories of stylometric features (i.e., lexical + syntactic + structural) outperforms the other combinations (i.e., (a) lexical + structural, (b) syntactic + structural, and (c) lexical + syntactic). These results indicate that the stylometric information captured by different features categories is complementary and orthogonal. Consequently, combining all feature categories improves the performance of authorship identification process; and (3) using partial Hausdorff distance, that is associated with outlier handling mechanism, outperforms the standard Hausdorff distance by 10.9 percentage points. This paper has laid the foundation for future work in Thai authorship identification task. We hope that this investigation has opened the door for future work on Thai to keep up with the work in other languages.

Acknowledgments

The research was partially supported by the Digital Economy Promotion Agency (project# MP-62-0003); and Thailand Research Fund and Office of the Higher Education Commission (MRG6180266).

References

[1] Malik H. Altakrori, Farkhund Iqbal, Benjamin C. M. Fung, Steven H. H. Ding, and Abdallah Tubaishat. 2019. Arabic Authorship Attribution: An Extensive Study on Twitter Posts. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.* 18, 1 (2019), 5:1–5:51.

[2] Sophia Ananiadou, Paul Thompson, and Raheel Nawaz. 2013. Enhancing search: Events and their discourse context. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 318–334.

[3] Riza Theresa Batista-Navarro, Georgios Kontonatsios, Claudiu Mihăilă, Paul Thompson, Rafal Rak, Raheel Nawaz, Ioannis Korkontzelos, and Sophia Ananiadou. 2013. Facilitating the analysis of discourse phenomena in an interoperable NLP platform. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 559–571.

[4] Malcolm Coulthard. 2012. On admissible linguistic evidence. *JL & Pol'y* 21 (2012), 441.

[5] Boonyarit Deewattananon and Usa Sammapun. 2017. Analyzing user reviews in Thai language toward aspects in mobile applications. In *14th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. 1–6.

[6] Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2019. NOVA: A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.* 18, 2 (2019), 17:1–17:18.

[7] Heba El-Fiqi, Eleni Petraki, and Hussein A. Abbass. 2016. Pairwise Comparative Classification for Translator Stylometric Analysis. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.* 16, 1 (2016), 2:1–2:26.

[8] Zhenhao Ge, Yufang Sun, and Mark J. T. Smith. 2016. Authorship Attribution Using a Neural Network Language Model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. 4212–4213.

[9] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. Deep learning. vol. 1.

[10] Saeed-Ul Hassan, Raheem Sarwar, and Amina Muazzam. 2016. Tapping into intra-and international collaborations of the Organization of Islamic Cooperation states across science and technology disciplines. *Science and Public Policy* 43, 5 (2016), 690–701.

[11] C.C. Holmes and N.M. Adams. 2002. A probabilistic nearest neighbour method for statistical pattern recognition. *J R Stat Soc Series B Stat Methodol* 64, 2 (2002), 295–306.

[12] Daniel P. Huttenlocher, Gregory A. Klanderman, and William Rucklidge. 1993. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* 15, 9 (1993), 850–863.

[13] Maham Jahangir, Hammad Afzal, Mehreen Ahmed, Khawar Khurshid, and Raheel Nawaz. 2017. An expert system for diabetes prediction using auto tuned multi-layer perceptron. In *2017 Intelligent Systems Conference (IntelliSys)*. IEEE, 722–728.

[14] Patrick Juola, George K. Mikros, and Sean Vinsick. 2019. A comparative assessment of the difficulty of authorship attribution in Greek and in English. *JASIST* 70, 1 (2019), 61–70.

[15] Natthawut Kertkeidkachorn, Proadpran Punyabukkana, and Atiwong Suchato. 2016. Acoustic Features for Hidden Conditional Random Fields-Based Thai Tone Classification. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.* 15, 2 (2016), 9:1–9:26.

[16] Nongnuch Ketui, Thanaruk Theeramunkong, and Chutamanee Onsuwan. 2015. An EDU-Based Approach for Thai Multi-Document Summarization and Its Application. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.* 14, 1 (2015), 4:1–4:26.

[17] Rajalida Lipikorn, Akinobu Shimizu, and Hidefumi Kobatake. 1994. A modified Hausdorff distance for object matching. In *Pattern Recognition*, Vol. 1. 566–568.

[18] Rangsipan Marukatat, Robroo Somkiadcharoen, Ratthanan Nalintasnai, and Tappasarn Aramboonpong. 2014. Authorship attribution analysis of thai online messages. In *IEEE International Conference on Information Science & Applications (ICISA)*. 1–4.

[19] Thomas M. Mitchell. 1997. *Machine Learning* (1 ed.). McGraw-Hill, Inc., New York, NY, USA.

[20] Frederick Mosteller and David Wallace. 1964. Inference and disputed authorship: The Federalist. *Reading MA: Addison-Wesley* (1964).

[21] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the feasibility of internet-scale author identification. In *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 300–314.

[22] Raheel Nawaz, Paul Thompson, and Sophia Ananiadou. 2012. Identification of Manner in Bio-Events.. In *LREC*. 3505–3510.

[23] Sarana Nutanong, Chenyun Yu, Raheem Sarwar, Peter Xu, and Dickson Chow. 2016. A Scalable Framework for Stylometric Analysis Query Processing. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*. 1125–1130.

[24] Mathias Payer, Ling Huang, Neil Zhenqiang Gong, Kevin Borgolte, and Mario Frank. 2015. What You Submit Is Who You Are: A Multimodal Approach for Deanonymizing Scientific Publications. *IEEE Trans. Information Forensics and Security* 10, 1 (2015), 200–212.

[25] Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. 2003. Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 267–274.

[26] Fahad Sabah, Saeed-Ul Hassan, Amina Muazzam, Sehrish Iqbal, Saira Hanif Soroya, and Raheem Sarwar. 2019. Scientific collaboration networks in Pakistan and their impact on institutional research performance. *Library Hi Tech* 37, 1 (2019), 19–29.

[27] Ahmad Al Sallab, Ramy Baly, Hazem M. Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro. 2017. AROMA: A Recursive Deep Learning Model for Opinion Mining in Arabic as a Low Resource Language. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.* 16, 4 (2017), 25:1–25:20.

[28] Raheem Sarwar and Saeed-Ul Hassan. 2015. A bibliometric assessment of scientific productivity and international collaboration of the Islamic World in science and technology (S&T) areas. *Scientometrics* 105, 2 (2015), 1059–1077.

[29] Raheem Sarwar, Qing Li, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. A Scalable Framework for Cross-lingual Authorship Identification. *Information Sciences* (2018).

[30] Raheem Sarwar and Sarana Nutanong. 2016. The Key Factors and Their Influence in Authorship Attribution. *Research in Computing Science* 110 (2016), 139–150.

[31] Raheem Sarwar, Saira Hanif Soroya, Amina Muazzam, Fahad Sabah, Sehrish Iqbal, and Saeed-Ul Hassan. 2019. A Bibliometric Perspective on Technology-Driven Innovation in the Gulf Cooperation Council (GCC) Countries in Relation to Its Transformative Impact on International Business. In *Technology-Driven Innovation in Gulf Cooperation Council (GCC) Countries: Emerging Research and Opportunities*. IGI Global, 49–66.

[32] Raheem Sarwar, Chenyun Yu, Sarana Nutanong, Norawit Urailertprasert, Nattapol Vannaboot, and Thanawin Rakthanmanon. 2018. A Scalable Framework for Stylometric Analysis of Multi-author Documents. In *Database Systems for Advanced Applications - 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24, 2018, Proceedings, Part I*. 813–829.

[33] Raheem Sarwar, Chenyun Yu, Ninad Tungare, Kanatip Chitavisutthivong, Sukrit Sriratanawilai, Yaohai Xu, Dickson Chow, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. An Effective and Scalable Framework for Authorship Attribution Query Processing. *IEEE Access* 6 (2018), 50030–50048.

[34] Fabrizio Sebastiani. 2006. Classification of text, automatic. *The encyclopedia of language and linguistics* 14 (2006), 457–462.

[35] Matthew Shardlow, Riza Batista-Navarro, Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2018. Identification of research hypotheses and new knowledge from scientific literature. *BMC medical informatics and decision making* 18, 1 (2018), 46.

[36] Thamar Solorio, Paolo Rosso, Manuel Montes-y-Gómez, Prasha Shrestha, Sebastián Sierra, and Fabio A. González. 2017. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. 669–674.

[37] Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2017. Enriching news events with meta-knowledge information. *Language Resources and Evaluation* 51, 2 (2017), 409–438.

[38] Ying Zhao and Justin Zobel. 2007. Searching With Style: Authorship Attribution in Classic Literature. In *Computer Science 2007. Proceedings of the Thirtieth Australasian Computer Science Conference (ACSC2007). Ballarat, Victoria, Australia, January 30 - February 2, 2007. Proceedings*. 59–68.